LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-478794

CBM.DIAGB.03.10.LLNL.007 Final Report

T. Slezak, M. Torres

April 5, 2011

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-TR-478794

CBM.DIAGB.03.10.LLNL.007 Final Report
Genomic Sequence Threat Characterization Pipeline

Thomas R. Slezak & Marisa W. Torres
Lawrence Livermore National Laboratory

3/29/2011

Executive Summary

Purpose of Project

The purpose of this project was to construct a system for characterizing the threat potential of genomic sequences, specifically assembled draft genomes. New genomes are characterized by initially comparing them against already-sequenced genomes. If the new genome is determined to be from a high-threat species, detailed (forensic-level) characterization is done based on gene and SNP (Single Nucleotide Polymorphism) data comparisons with all other previously-sequenced members of that high-threat species. New genomes are compared against a large set of known virulence and antibiotic-resistance genes and also compared against a large set of vectors that could be used for bacterial genetic engineering. Together, these analyses provide a comprehensive initial assessment of the most likely phylogenetic placement of a new genome, plus an assessment of the known-gene content and an indication of any possible bacterial genetic engineering utilizing vector-mediated techniques. This provides an initial threat potential summary based on high information content comparisons (e.g., thousands of genes, SNPs, and potential genetic engineering vectors) that can be used to guide subsequent operational response or more detailed laboratory characterization.

Project Completion

All five analysis modules (Species Determination, SNP Analysis, Family Gene Analysis, Virulence/Resistance Analysis, Genetic Engineering Vector Analysis) have been completed and integrated into the threat characterization pipeline. Information from these analysis modules has been integrated into the Microbial Forensic Encyclopedia (MFE), leveraging database infrastructure built under DHS funding. This allowed resources to be concentrated on analyses instead of having to build duplicate infrastructure.

Twelve genomes were processed through the threat characterization pipeline. The analysis was across four milestone categories: known finished and draft BWA genomes, bacterial vector, and blinded input. See the Milestone Results section for summary details.

Milestone1 - Characterize known finished BWA genomes

- Francisella tularensis SCHU S4 was correctly characterized by species and SNP analysis.
- Variola virus India 67 was correctly characterized by species and SNP analysis.

Milestone 2 - Characterize known draft BWA genomes

- Bacillus anthracis A0248 was correctly characterized by SNP analysis that indicated a match to BAB which is A0248 isolate B.
- Brucella melitensis ATCC 23457 was correctly characterized by species, SNP, family gene, and virulence analysis.
- Burkholderia mallei NCTC 10247 was correctly characterized by species, SNP, family gene, and virulence analysis.
- Burkholderia pseudomallei 668 was correctly characterized by species and family gene analysis.

- Francisella tularensis Wy96 3418 was correctly characterized by species, family gene, and virulence analysis.
- Yersinia pestis Harbin 35 was correctly characterized by species, SNP, family gene, and virulence analysis.
- Ebola virus Sudan Boniface was correctly characterized by species and family gene analysis. Ebola virus Zaire Mayinga was also characterized by species, family gene, and virulence analysis.

Milestone3 - Characterize bacterial vector data

- pBAD18 cloning vector was characterized as a cloning vector by species and also having evidence of a genetic engineering vector. The pBAD18 cloning vector was considered the "clean" test case, with only the test vector present.
- Vector simulants using varying amounts of vector in host bacteria genome were characterized. See the Discussion section for assessment of the simulant results.

Milestone 4 - Characterize blinded input genomes from TMT

- GPSG4HP01 was characterized as Klebsiella pneumoniae by species, family gene, and virulence analysis and having evidence of a genetic engineering vector.
- GPV77IU02 was characterized as Klebsiella pneumoniae by species and family gene analysis and having evidence of a genetic engineering vector.

This one-year project is completed. The remaining effort was focused on additional testing and completing a report generator that automatically summarized the most important results from each of the 5 analysis modules. This report updates the summary from the project's annual report. A large .zip file attachment contains the detailed outputs from each of the milestone tests listed above.

In mid-February Dr. Nicole Rosenzweig of ECBC visited LLNL and was briefed on the threat characterization pipeline. She leads the TMT effort to build sequence characterization infrastructure, so we discussed with her how this project might aid the TMT effort.

Background

Historically, most existing genome sequence analysis systems have focused on *de novo* gene-finding. Examination of Genbank entries for sequenced bacteria show that a large percentage of the genes found by such systems are "hypothetical, unknown function." While useful for guiding basic research projects on determining gene function, these analyses are of relatively little use in operational scenarios of relevance to DTRA. We are focusing instead on genome sequence analyses that provide a threat characterization perspective.

The particular scenario we have focused on for this project is analyzing a just-sequenced draft genome from a sample of interest to DTRA. Our approach has been focused on answering several basic questions about such a just-sequenced draft genome:

1. What organism(s) are most likely present in this assembled draft genome?
2. If any high-threat organism(s) are present, what are their most likely closest already-sequenced relatives?
3. What known virulence and resistance mechanisms appear to be present (or missing)?
4. Is there any indication of potential vector-mediated bacterial genetic engineering?

Question (1) acknowledges that a sample may contain a chimeric organism (e.g., consists of portions of more than one known species. This could occur either naturally or via deliberate genetic engineering). Alternatively, the sample could contain a contaminant or might not be as "pure" as originally thought. We note that this does not imply that we are handling the full metagenomic sequence analysis problem (since that means analyzing unassembled sequence reads instead of assembled contigs.) An original 2nd year planned for this project that would have dealt with metagenomic sequence analysis was cancelled.

Question (2) applies if it was determined in (1) that a high-threat genome is likely present. The Category A bacteria are key examples. In these cases, it is desirable to use phylogenetic analysis methods to place the new genome properly against already-sequenced genomes from the same high-threat species, using the highest-resolution methods possible. We perform this analysis using 2 orthogonal techniques: SNPs and gene presence/absence. We have determined thousands of single nucleotide polymorphisms that can be utilized to classify the evolutionary relationship between different strains of the same species to very high resolution. Additionally, we can use the presence/absence status of a large class of genes (all non-redundant genes from Genbank RefSeq genomes of the high-threat species and all near-neighbors) to provide an independent phylogenetic assignment of the new genome among already-sequenced ones. It is worth noting that the SNP analysis provides greater resolution (single nucleotide) but in most cases there is no functional significance known about the variations. The gene presence/absence analysis inherently has a much lower resolution; however it can provide a measure of functional knowledge (for genes for which function is purportedly known.) For example, a certain SNP value may imply that the unknown genome is closest to a particular strain, but the gene presence/absence analysis might indicate whether the unknown genome is likely "fully loaded" or whether some important genes might be missing. We have the ability to leverage both SNP and gene analyses originally performed for NBACC to provide the data for this analysis.

Question (3) is primarily focused on potential payloads of bacterial genetic engineering. Leveraging other prior DHS and DTRA funded work, we compare a large set of known virulence and antibiotic-resistance genes against the new draft genome. In some cases the results may indicate obvious issues to examine further (e.g., the anthrax toxin gene should not be present in a *Salmonella* genome) but others would require high-level analyses that are outside of the scope of this project (e.g., is vancomycin resistance expected to be found in *Clostridium perfringens* isolated from a wound sample?) Note that the virulence and resistance gene information could prove useful to guide countermeasures. However, the apparent presence of a gene does not prove that it is actually expressed; follow-up laboratory experiments would be needed to confirm any predictions about particular virulence and resistance potential.

Question (4) focuses on determining if significant portions of any known vectors or plasmids appear to be present in the draft genome being examined. Such "vector scars" could be indications of deliberate vector-mediated bacterial genetic engineering. Leveraging a collection of vector/plasmid sequences originally developed under IC funding, we can compare it against the draft genome to look for significant hits. Any substantial hits can be considered to raise a yellow flag warning to check further. If coupled with any unexpected gene presence from (3) above, further lab testing would certainly be warranted.

Answering these questions about a draft genome will provide rapid, high-confidence initial actionable information for both response and attribution.

Discussion

Examination of the gene family and virulence phylogenetic trees in the annual report's Milestone Result section highlighted the need to adjust the data threshold filters. This adjustment was performed and shifted the query sequence into the expected groupings. Corrected phylogenetic trees are included in the final report. There were non-query "unidentified" sequences that were showing up in some phylogenetic trees from other pipeline run test in the annual report's results. The final report phylogenetic trees include only a single unknown pipeline sequence.

To better understand the sensitivity of the vector search results, we ran a series of tests on sequences with varying amounts of specific known vector sequences inserted. While the vector search results appear to be reasonably complete when large amounts of the inserted vector sequence are present, the picture quickly becomes confusing as the amount of vector sequence decreases. Our interpretation of this behavior, is that more development is needed on the genetic engineering vector search tools, and on annotation of the vector databases.

The results of our vector sequence sensitivity tests show that the vector probe results are best for identifying potential vector sequences when present at low levels. When a large amount of vector sequence is present, both the probe results and the BLAST results can provide identification.

In the cases where we inserted large contiguous sections (of approximately 2-4 kb) of our vector simulant sequences, the simulant sequences that we inserted appear prominently in both the

vector probe results, and in the vector BLAST results. However, when the amount of vector sequence is only 500bp, interpretation of the results is much more complicated.

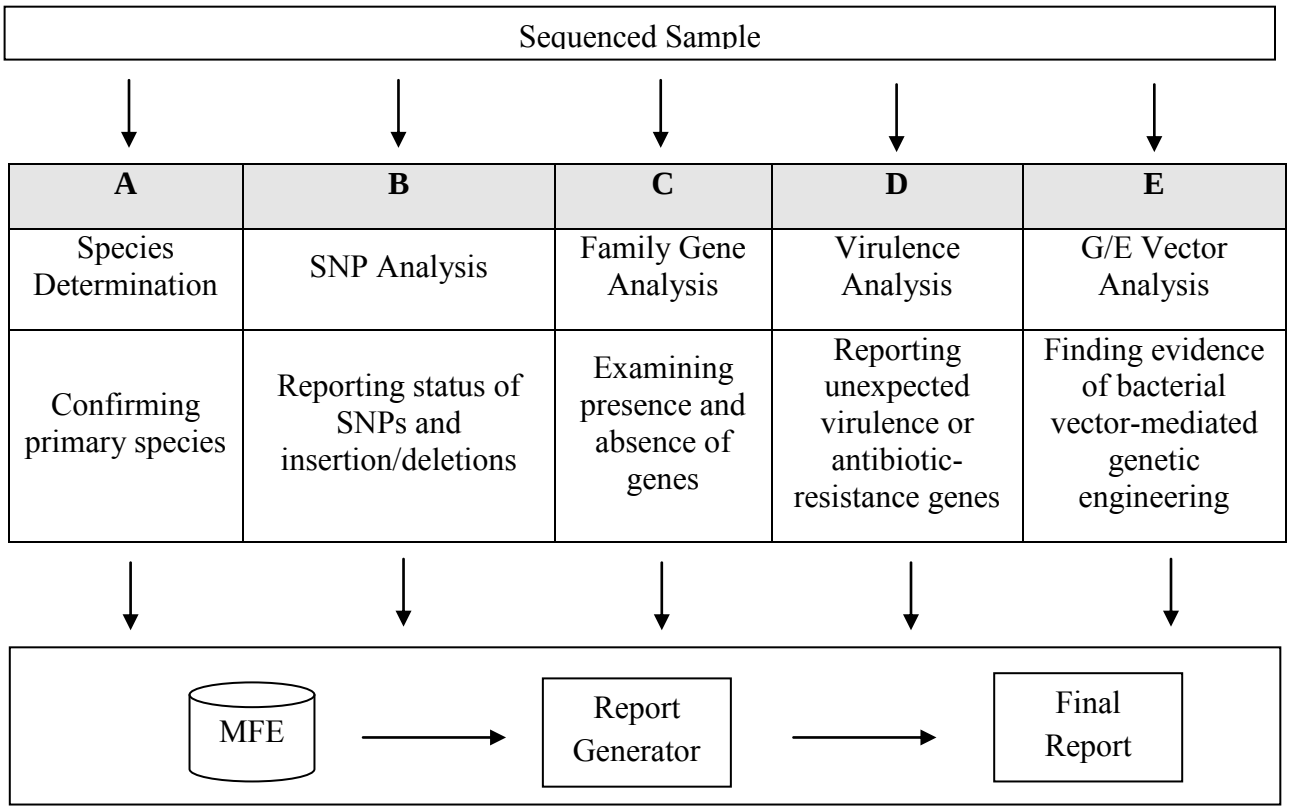
The vectors appearing in the blast alignment lists correlate less and less with the probe results, as the amount of vector sequence decreases. In many cases this could be misleading, since the coverage of the vectors is not, on its own, an accurate measure. The BLAST results are instead helpful when trying to identify the amount of the vector sequence that was present in the sample (the coverage percent).

Many of the GE vector titles aren't informative enough to be intelligently grouped by type. Even in the case of a vector mixture, where 5k of the main "spiked" vector sequence is present, and 2k of a secondary vector sequence is present, the secondary is not present in the top 10 blast results by coverage percent (range 92.97-99.98%) or length (range 4,174-11,263bp). The secondary sequence was outside of the top 10 range with a coverage percent of 57.3% and a coverage length of 2,687bp, so this secondary vector sequence was buried in the results list.

This highlights that the primary use of the BLAST results is as an additional confirmation of probe results, since it provides additional coverage information.

Pipeline Diagram

A sequenced sample is processed through five analysis pathways (A-E). The appropriate run parameters for the analysis are automatically determined to handle processing unknowns. The data from these pathways is integrated into the MFE (Microbial Forensics Encyclopedia) database. We completed the report generator tool, which prepares the data into a final report.



Overall System Architecture

The operation of each analysis module is described below:

- Species Determination is accomplished via MUMmer analysis against our comprehensive database of finished and draft microbial genomes. Each contig is compared separately (since the sample sequenced could be chimeric or contain multiple organisms). An aggregation of the top hits are used to determine the closest matches, based on raw genomic sequence matching.
- SNP analysis is performed if the results of the Species Determination module indicate a strong likelihood that a key threat agent species may be present. We have comprehensive SNP data for the category A bacterial agents (BA, YP, FT, Brucella, Burkholderia mallei, Orthopox virus). The new draft genome is compared against the set of SNPs (note that this is on the order of at least 4,000 SNPs, determined by analysis of all available genomes for these key species.) A SNP value vector is created for the new genome, and run against all data for available genomes of that species in a phylogenetic analysis. This

results in the new genome being placed appropriately to its nearest neighbors (based on SNP-encoded evolutionary distance.) This forensic placement may be of use to determine where a sample may have come from, and whether or not its appearance in a particular location requires deeper investigation.

- Family Gene Analysis is also performed if the results of the Species Determination module indicate a strong likelihood that a key threat agent species may be present. We have created a non-redundant gene list at the taxonomic Family level for each of the category A bacterial agents listed above. This was created by extracting each RefSeq gene from Genbank associated with all species in the same taxonomic Family as the key threat agents. Genes were considered redundant if there was > 90% nucleotide similarity over > 90% of the gene length. We compare this large (on the order of ~10,000) set of genes against the new draft genome and create a presence/absence vector using the same 90% similarity over 90% gene length. The presence/absence vector of the new draft genome is compared with the vectors from all other sequenced genomes from the threat agent Family taxonomic level in a phylogenetic analysis. This provides an orthogonal forensic assignment to that provided by the SNP analysis module. We note that the current use of a 90% similarity cutoff is arbitrary and phylogenetic placement may vary if stricter or looser values are used.
- We perform a Virulence Analysis for all draft genomes. We compare via BLAST the 23,029 virulence-associated and 5,625 antibiotic-resistance proteins from our MvirDB (Microbial Virulence DataBase) against the draft genome. Hits above a threshold, 90% similarity over 90% protein length, set high enough to be confident that a highly-similar gene is present, are recorded. We note that at this time, the TMT project is considering how to perform decision support from this kind of gene presence and absence information. That is a separate project and although we can provide the input to such decision support, it is beyond the scope of this one-year project. We also note that our original plan to also test the 60bp virulence microarray probes LLNL developed for the IC via BLAST proved inferior to using BLAST on the entire virulence protein sequences. As noted in our earlier progress reports, we discarded checking the microarray probes.
- We perform a Genetic Engineering Vector Analysis for all draft genomes. We have a set of about 3,800 vectors and plasmids whose presence in a draft genome might be indicative of potential deliberate bacterial genetic engineering. We compare these vectors and plasmids using BLAST against the draft genome and note any significant hits. This approach can flag via detected "vector scars" in the draft genome that some sort of vector-mediated bacterial gene insertion may have been performed. It acts as a potential "yellow flag" for closer inspection. Since most vectors and plasmids used for bacterial engineering have natural origins, some manual interpretations of the results are required. For example, many vectors have a modified *E. coli* backbone, and thus may indicate spurious potential "hits" in *E. coli* or *Shigella* strains. We also test the 60bp genetic engineering microarray probes LLNL developed for NBACC via BLAST. The vector probe results are important, since the probes were designed on functional vector regions that are unique from naturally occurring plasmids.

The 5 Analysis Modules are currently run from a master Python program. Our development has taken place on a multi-cpu Sun Solaris server, however, all codes should be portable to common versions of Linux on commodity Intel/AMD hardware. We note that there are ample

opportunities to improve run time on particular implementation instances. Utilizing cluster nodes efficiently in parallel requires converting to whatever specific job scheduler is being supported on a target cluster computer. Since these are highly non-standard, we have not optimized for overall analysis run time in this one-year project.

The report generator queries the highest ranked results from the MFE database and prepares the final report tables as described below:

- Ranking Summary table overlays the ranked results from all the analysis types with the NCBI taxonomy tree. This facilitates comparison across multiple result types for making an overall assessment. A sequence match with multiple high rankings would give higher confidence in the supporting evidence.
- Species Determination tables report the probe level species call and the mummer alignment full genome confirmation.
- SNP Analysis table links to the SNP phylogenetic tree.
- Family Gene Analysis tables report the highest ranked sequences by gene content similarity. Each chromosome or plasmid is reported separately. There is a link to the family gene phylogenetic tree.
- Virulence Analysis tables reports the highest ranked sequences by virulence similarity. Each chromosome or plasmid is reported separately. There is a link to the virulence phylogenetic tree. The report includes the count of genes by virulence category.
- G/E Vector Analysis tables report the probe level G/E vector call and the blast comparison of full G/E vector sequence confirmation. Two representations of genetic engineering vector results are presented in the each of the individual sequence reports. First is the results of the vector probe search, expressed as log-odds, just like the species determination results. The second representation is BLAST coverage, which is sometimes useful additional information to complement the vector probe results.

Milestone Results Summary

Due to the volume of report results, the accompanying .zip file contains the results referenced below with specific .docx file names.

M1 - Characterize known finished BWA genomes

Francisella tularensis SCHU S4

Results in Francisella_tularensis_SCHU_S4/results.docx. A, B, and E's results correctly characterized the query as close to Francisella tularensis SCHU S4 with no genetic engineering evidence. C and D's results both had SCHU S4 as the second highest match.

Variola virus India 67

Results in Variola_virus_India_67/results.docx. A, B, and E's results correctly characterized the query as Variola virus India 67 with no genetic engineering evidence.

M2 - Characterize known draft BWA genomes

Bacillus anthracis A0248

Results in Bacillus_anthracis_A0248/results.docx. B and E's results correctly characterized the query as Bacillus anthracis A0248 with no genetic engineering evidence. A, C, and D's results indicated a similarity with anthracis sequences.

Brucella melitensis ATCC 23457

Results in Brucella_melitensis_ATCC_23457/results.docx. A, B, C, D, and E's results correctly characterized the query as Brucella melitensis ATCC 23457 with no genetic engineering evidence.

Burkholderia mallei NCTC 10247

Results in Burkholderia_mallei_NCTC_10247/results.docx. A, B, C, D, and E's results correctly characterized the query as Burkholderia mallei NCTC 10247 with no genetic engineering evidence.

Burkholderia pseudomallei 668

Results in Burkholderia_pseudomallei_668/results.docx. A, C, and E's results correctly characterized the query as Burkholderia pseudomallei 668 with no genetic engineering evidence. D's results had 668 as the third highest match.

Francisella tularensis Wy96 3418

Results in Francisella_tularensis_Wy96_3418/results.docx. A, C, D, and E's results correctly characterized the query as Francisella tularensis Wy96 3418 with no genetic engineering evidence. B's results had a close match to Wy96 3418.

Yersinia pestis Harbin 35

Results in Yersinia_pestis_Harbin_35/results.docx. A, B, C, D, and E's results correctly characterized the query as Yersinia pestis YPC Harbin 35 with no genetic engineering evidence.

Ebola virus Sudan Boniface

Results in Ebola_virus_Sudan_Boniface/results.docx. A, C, and E's results correctly characterized the query as Ebola virus Sudan Boniface with no genetic engineering evidence. A, C, and D also characterized the query as Ebola virus Zaire Mayinga. D's virulence placement to only Ebola Zaire suggests that virulence factors were not defined for the Sudan clade.

M3 - Characterize bacterial vector data

pBAD18 cloning vector

Results in pBAD18_cloning_vector/results.docx. A, B, and E's results characterized the query as a cloning vector with genetic engineering evidence. We note that we do not have access to any sequenced genomes that have had deliberate bacterial vector-mediated genetic engineering. As a substitute, we can run vectors through the pipeline. Many vectors are quite similar, hence we do not expect to get perfect matches, just an indication that some vector appears to be present.

M4 - Characterize blinded input genomes from TMT

GPSG4HP01

Results in GPSG4HP01/results.docx. A, C, D and E's results characterized the query as Klebsiella pneumoniae with genetic engineering evidence. TMT can supply the exact identity of this blinded Exercise 2 sample.

GPV77IU02

Results in GPV77IU02/results.docx. A, C, and E's results characterized the query as Klebsiella pneumoniae with genetic engineering evidence. D's results did not indicate a close similarity to Klebsiella pneumoniae sequences. TMT can supply the exact identity of this blinded Exercise 2 sample.

Detailed Report Examples

The full reports are in an accompanying 11MB .zip file. We have included samples here of what information is contained for each of the sequence analyses performed using the pipeline.

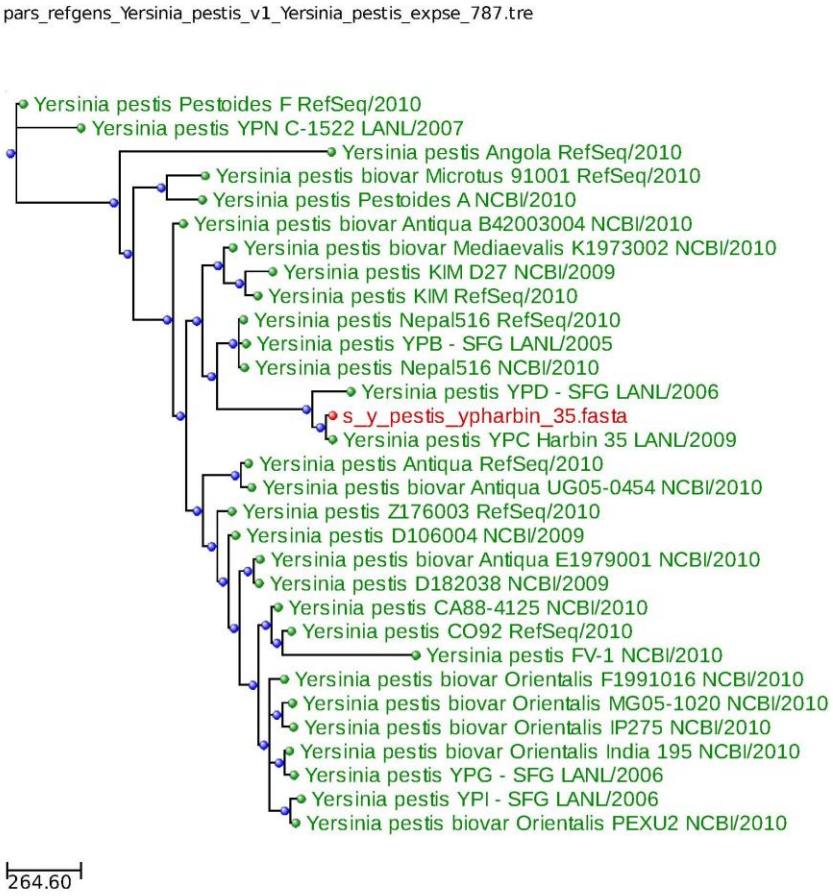


Figure 1. SNP phylogenetic tree showing placement of the YP Harbin genome (in red)

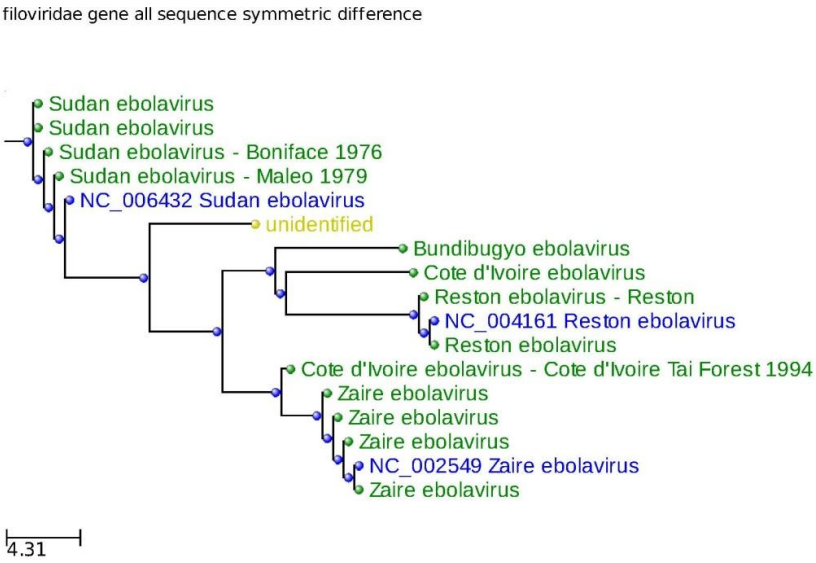


Figure 2. Phylogenetic tree based on gene content for an Ebola Sudan Boniface genome (listed as "unidentified" in yellow). Viruses have few genes, compared to bacteria. See the .zip files for the very large high-resolution gene content phylogenetic trees for the bacterial genomes processed.

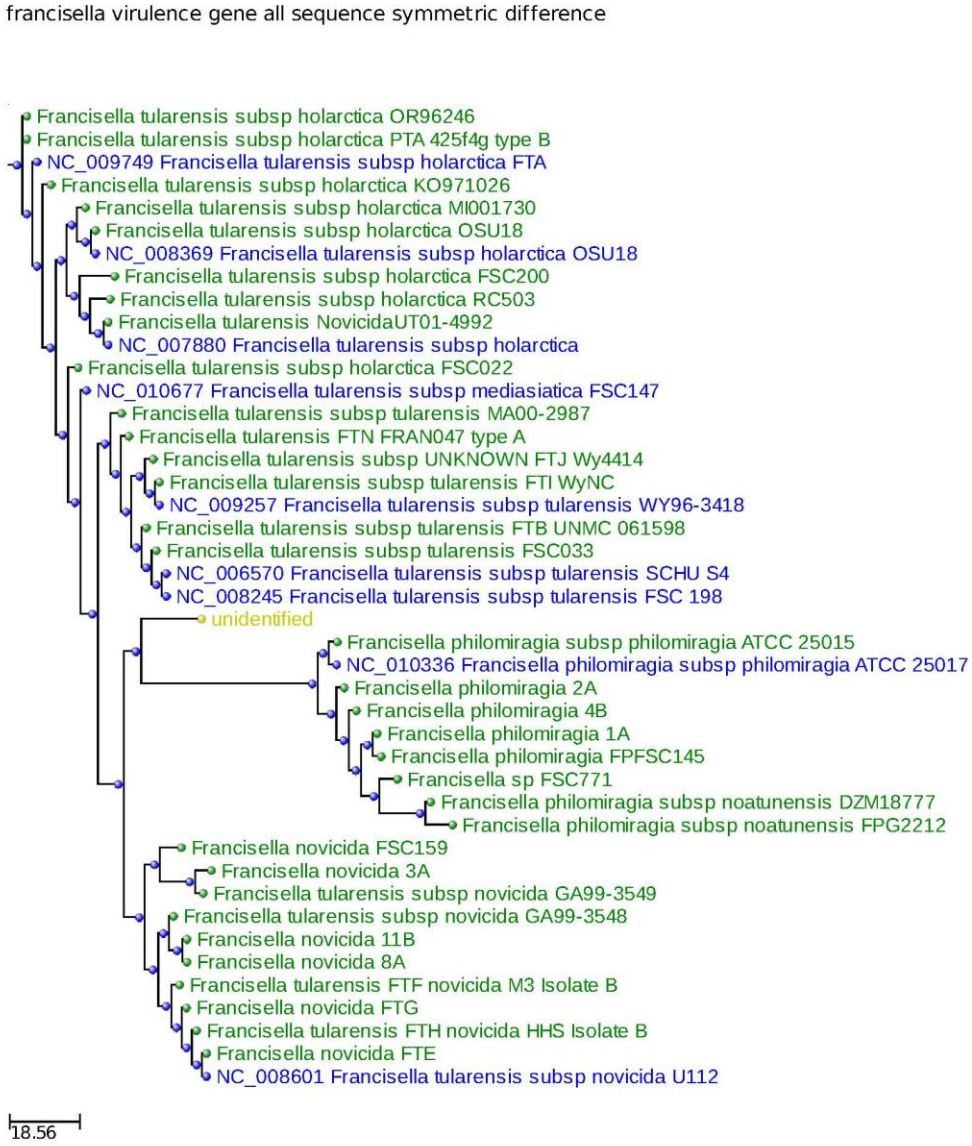
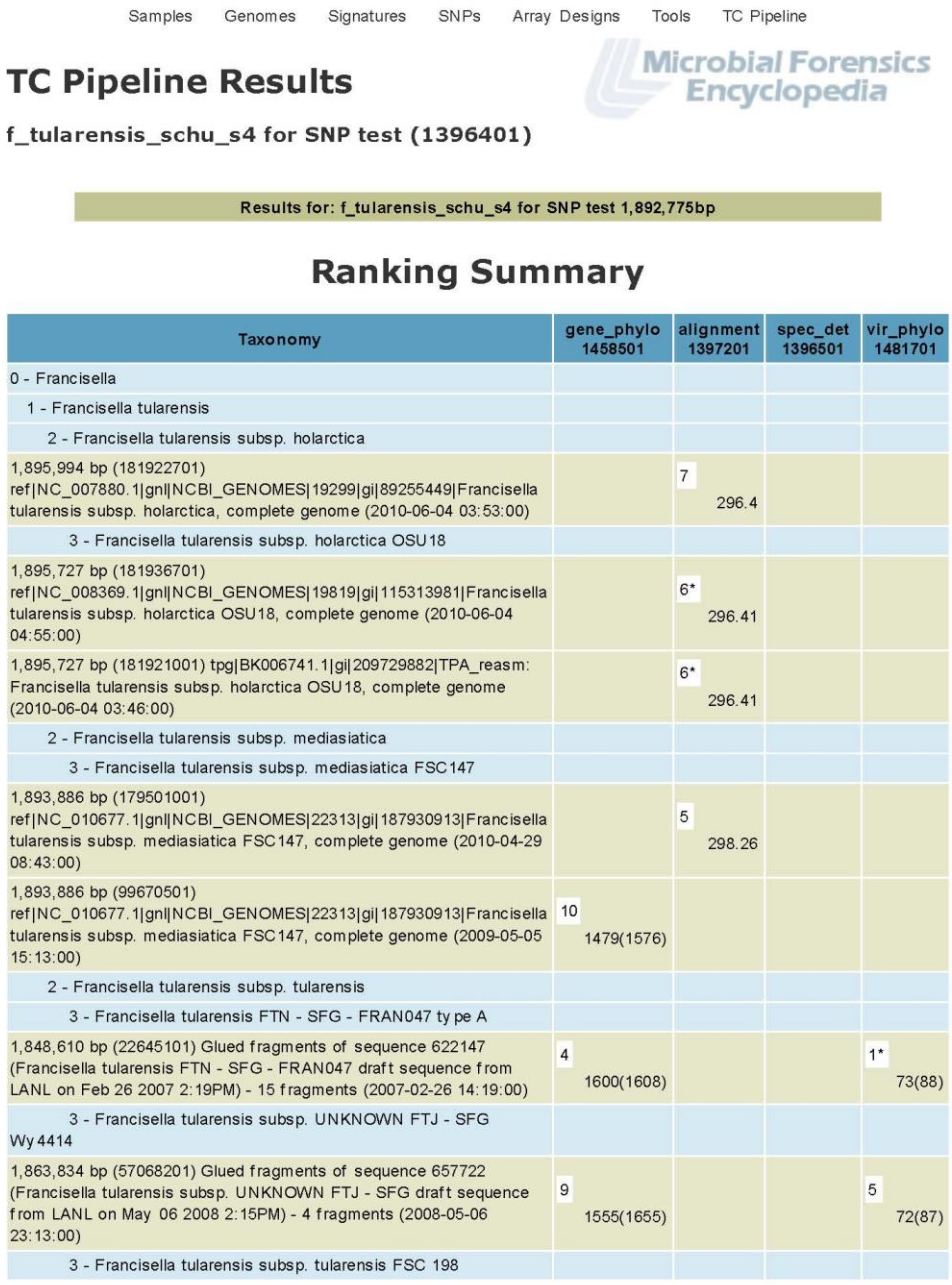


Figure 3. Phylogenetic tree based on virulence and resistance gene content (only) for a FT SCHU 4 genome (shown as "unidentified" in yellow). This is the 3rd independent, orthogonal phylogenetic placement approach implemented in the pipeline (e.g., SNPs, family total gene content, and known virulence/resistance gene content.) Note that this virulence/resistance method would flag any genetic engineering into benign organisms.



Samples Genomes Signatures SNPs Array Designs Tools TC Pipeline

Microbial Forensics Encyclopedia

TC Pipeline Results

f_tularensis_schu_s4 for SNP test (1396401)

Results for: f_tularensis_schu_s4 for SNP test 1,892,775bp

Ranking Summary

Taxonomy	gene_phylo 1458501	alignment 1397201	spec_det 1396501	vir_phylo 1481701
0 - Francisella				
1 - Francisella tularensis				
2 - Francisella tularensis subsp. holarctica				
1,895,994 bp (181922701) ref|NC_007880.1|gnl|NCBI_GENOMES|19299|gi|89255449|Francisella tularensis subsp. holarctica, complete genome (2010-06-04 03:53:00)		7 296.4		
3 - Francisella tularensis subsp. holarctica OSU18				
1,895,727 bp (181936701) ref|NC_008369.1|gnl|NCBI_GENOMES|19819|gi|115313981|Francisella tularensis subsp. holarctica OSU18, complete genome (2010-06-04 04:55:00)		6* 296.41		
1,895,727 bp (181921001) tpg|BK006741.1|gi|209729882|TPA_reasm: Francisella tularensis subsp. holarctica OSU18, complete genome (2010-06-04 03:46:00)		6* 296.41		
2 - Francisella tularensis subsp. mediasiatica				
3 - Francisella tularensis subsp. mediasiatica FSC147				
1,893,886 bp (179501001) ref|NC_010677.1|gnl|NCBI_GENOMES|22313|gi|187930913|Francisella tularensis subsp. mediasiatica FSC147, complete genome (2010-04-29 08:43:00)		5 298.26		
1,893,886 bp (99670501) ref|NC_010677.1|gnl|NCBI_GENOMES|22313|gi|187930913|Francisella tularensis subsp. mediasiatica FSC147, complete genome (2009-05-05 15:13:00)	10 1479(1576)			
2 - Francisella tularensis subsp. tularensis				
3 - Francisella tularensis FTN - SFG - FRAN047 type A				
1,848,610 bp (22645101) Glued fragments of sequence 622147 (Francisella tularensis FTN - SFG - FRAN047 draft sequence from LANL on Feb 26 2007 2:19PM) - 15 fragments (2007-02-26 14:19:00)	4 1600(1608)			1* 73(88)
3 - Francisella tularensis subsp. UNKNOWN FTJ - SFG Wy4414				
1,863,834 bp (57068201) Glued fragments of sequence 657722 (Francisella tularensis subsp. UNKNOWN FTJ - SFG draft sequence from LANL on May 06 2008 2:15PM) - 4 fragments (2008-05-06 23:13:00)	9 1555(1655)			5 72(87)
3 - Francisella tularensis subsp. tularensis FSC 198				

Figure 4 (includes 7 pages below as well). This is the integrated pipeline report that is automatically produced within the context of the DHS-funded Microbial Forensic Encyclopedia. The FT SCHU4 example is shown. Note the sections for each of the analysis modes, including summaries, some raw data, and URLs to the phylogenetic tree PDFs (including more variants than were shown in examples 1-3.)

1,892,616 bp (181936001) ref|NC_008245.1|gnl|NCBI_GENOMES|19668|gi|110669657|Francisella tularensis subsp. tularensis FSC198, complete genome (2010-06-04 04:52:00)		2 299.99		
1,892,616 bp (99841801) ref|NC_008245.1|gnl|NCBI_GENOMES|19668|gi|110669657|Francisella tularensis subsp. tularensis FSC198, complete genome (2009-05-10 14:31:00)	1 1624(1632)			2* 73(89)
1,892,616 bp (47148001) ref|NC_008245.1|gnl|NCBI_GENOMES|19668|gi|110669657|Francisella tularensis subsp. tularensis FSC198, complete genome (2008-02-27 22:25:00)			1 229.6151	
3 - Francisella tularensis subsp. tularensis FSC033				
1,847,789 bp (91872201) Glued fragments of sequence 695074 (Francisella tularensis subsp. tularensis FSC033 Francisella tularensis subsp. tularensis FSC033, unfinished sequence, whole genome shotgun sequencing project from NCBI on...) - 15 fragments (2009-03-03 02:59:00)	5 1590(1603)			2* 73(89)
3 - Francisella tularensis subsp. tularensis FTB - SFG - UNMC 061598				
1,892,681 bp (56878401) Contig36 ./FTB.fasta.screen.ace.40 from 652 to 1893332|Francisella tularensis subsp. tularensis FTB - SFG - UNMC 061598 complete genome from LANL on May 06 2008 2:20PM (2008-05-06 14:20:00)	3 1622(1632)	3* 299.98		4* 73(89)
3 - Francisella tularensis subsp. tularensis FTI - SFG - Wy NC				
1,898,494 bp (56878301) Contig6 ./FTI.fasta.screen.ace.62 from 21 to 1898514|Francisella tularensis subsp. tularensis FTI - SFG - Wy NC draft sequence from LANL on May 06 2008 2:16PM (2008-05-06 14:20:00)	8 1565(1665)	4* 299.24		1* 73(88)
3 - Francisella tularensis subsp. tularensis MA00-2987				
1,876,942 bp (95005901) Glued fragments of sequence 695628 (Francisella tularensis subsp. tularensis MA00-2987 Francisella tularensis subsp. tularensis MA00-2987, unfinished sequence, whole genome shotgun sequencing project from N...) - 33 fragments (2009-03-09 20:35:00)	6 1535(1547)			
3 - Francisella tularensis subsp. tularensis NE061598				
1,892,681 bp (179365001) gnl|UNebraska|Seq1|gb|CP001633.1|gi|282158286|Francisella tularensis subsp. tularensis NE061598, complete genome (2010-04-29 04:52:00)		3* 299.98		
3 - Francisella tularensis subsp. tularensis SCHU S4				
1,892,775 bp (181917201) ref|NC_006570.2|gnl|NCBI_GENOMES|563|gi|255961454|Francisella tularensis subsp. tularensis SCHU S4, complete genome (2010-06-04 03:30:00)		1 300.0		
1,892,819 bp (99626001) ref|NC_006570.1|gnl|NCBI_GENOMES|563|gi|56707187|Francisella tularensis subsp. tularensis Schu 4, complete genome (2009-05-04 23:17:00)	2 1621(1629)			4* 73(89)
3 - Francisella tularensis subsp. tularensis WY96-3418				
1,898,476 bp (99846801) ref|NC_009257.1|gnl|NCBI_GENOMES|20753|gi|134301169|Francisella tularensis subsp. tularensis WY96-3418, complete genome (2009-05-10 15:14:00)	7 1568(1668)			1* 73(88)

1,898,476 bp (181940601) ref|NC_009257.1|gnl|NCBI_GENOMES|20753|gi|134301169|Francisella tularensis subsp. tularensis WY96-3418, complete genome (2010-06-04 05:12:00)		4* 299.24		
1 - Francisella tularensis subsp. novicida				
2 - Francisella novicida 11B				
1,873,974 bp (107614601) Glued fragments of sequence 706943 (Francisella novicida 11B draft sequence from USAMRIID on Jun 25 2009 2:24PM) - 94 fragments (2009-06-25 14:33:00)				6 71(86)
2 - Francisella tularensis subsp. novicida GA99-3548				
1,849,843 bp (91871701) Glued fragments of sequence 695069 (Francisella tularensis subsp. novicida GA99-3548 Francisella tularensis subsp. novicida GA99-3548, unfinished sequence, whole genome shotgun sequencing project from NCBI ...) - 18 fragments (2009-03-03 02:52:00)				3 72(87)
2 - Francisella tularensis subsp. novicida GA99-3549				
1,901,024 bp (42306201) Glued fragments of sequence 637775 (Francisella tularensis subsp. novicida GA99-3549 Francisella tularensis subsp. novicida GA99-3549, unfinished sequence, whole genome shotgun sequencing project from NCBI ...) - 15 fragments (2007-08-18 19:20:00)			2 8.2463	

A - Species Determination

Confirming primary species

A1 - Species Determination

Sequence/Contig Name	Length(bp)	Log Odds
MS_47148001 Francisella tularensis subsp tularensis FSC 198 Francisella	1,892,616	229.6151
MS_42306201 Francisella tularensis subsp novicida GA99-3549 Glued fragments of sequence 637775 Francisella tularensis subsp novicida GA99-3549 Francisella tularensis subsp novicida GA99-3549 unfinished sequence whole genome shotgun sequencing project	1,901,024	8.2463

A2 - Species Determination - Mummer Results (top 10)

Sequence/Contig Name	Length(bp)	AvgID%	Query%	Ref%
MR_5681701 Francisella tularensis subsp tularensis SCHU S4 Francisella	1,892,775	100.0	100.0	100.0
MR_5675201 Francisella tularensis subsp tularensis FSC 198 Francisella	1,892,616	99.99	100.0	100.0
MR_5675801 Francisella tularensis subsp tularensis FTB - SFG - UNMC 061598 Francisella	1,892,681	99.98	100.0	100.0
MR_5676401 Francisella tularensis subsp tularensis NE061598 Francisella	1,892,681	99.98	100.0	100.0
MR_5676201 Francisella tularensis subsp tularensis FTI - SFG - WyNC Francisella	1,898,494	99.69	99.82	99.73
MR_5682101 Francisella tularensis subsp tularensis WY96-3418 Francisella	1,898,476	99.69	99.82	99.73
MR_5673201 Francisella tularensis subsp mediasiatica FSC147 Francisella	1,893,886	99.44	99.29	99.53

MR_5672101 Francisella tularensis subsp holarctica OSU18 TPA_reasm	1,895,727	99.3	97.8	99.31
MR_5672201 Francisella tularensis subsp holarctica OSU18 Francisella	1,895,727	99.3	97.8	99.31
MR_5665501 Francisella tularensis subsp holarctica Francisella	1,895,994	99.32	97.76	99.32

B - SNP Analysis

Reporting status of SNPs and insertion/deletions

B.1 - SNP - Tritool Results		
Assay	**Reference Genome**	**Results Page**
Francisella	Francisella_tularensis	Snp Results Page
Francisella	Francisella_genus	Snp Results Page

C - Family Gene Analysis

Examining presence and absence of genes

C.1 - Family Gene Analysis

C.1.1 - Highest Ranked Sequences by Gene Content Similarity

Matched 1634 / 3370 total francisella genes

Sequence/Contig Name	Length(bp)	Ref Gene Count	Shared	Query Missing	Ref Missing
MR_4342101 Francisella tularensis subsp tularensis FSC 198 Francisella	1,892,616	1632	1624	8	10
MR_4342001 Francisella tularensis subsp tularensis SCHU S4 Francisella	1,892,819	1629	1621	8	13
MR_4342401 Francisella tularensis subsp tularensis FTB - SFG - UNMC 061598 Francisella	1,892,681	1632	1622	10	12
MR_4338201 Francisella tularensis FTN - SFG - FRAN047 type A Glued fragments of sequence 622147 Francisella tularensis FTN FRAN047 draft sequence	1,848,610	1608	1600	8	34
MR_4342201 Francisella tularensis subsp tularensis FSC033 Glued fragments of sequence 695074 Francisella tularensis subsp tularensis FSC033 Francisella tularensis subsp tularensis FSC033 unfinished sequence whole genome shotgun sequencing project	1,847,789	1603	1590	13	44

MR_4338101 Francisella tularensis subsp tularensis MA00-2987 Glued fragments of sequence 695628 Francisella tularensis subsp tularensis MA00-2987 Francisella tularensis subsp tularensis MA00-2987 unfinished sequence whole genome shotgun sequencing project	1,876,942	1547	1535	12	99
MR_4342301 Francisella tularensis subsp tularensis WY96-3418 Francisella	1,898,476	1668	1568	100	66
MR_4338001 Francisella tularensis subsp tularensis FTI - SFG - Wy NC Francisella	1,898,494	1665	1565	100	69
MR_4338301 Francisella tularensis subsp UNKNOWN FTJ - SFG Wy 4414 Glued fragments of sequence 657722 Francisella tularensis subsp UNKNOWN FTJ draft sequence	1,863,834	1655	1555	100	79
MR_4339501 Francisella tularensis subsp mediasiatica FSC147 Francisella	1,893,886	1576	1479	97	155

C.1.2 - Gene Significance					
Top Species	**Sequence Count**	**Gene Count**	**Shared**	**Query Missing**	**Species Missing**
Francisella tularensis	22	995	990	216	7

C.1.3 - Phylogenetic Tree	
Sequence Group Count	43
Sequence Group Type	all sequence
Pairwise Gene Comparison	symmetric difference
Phylogenetic Tree Pdf	
Newick Readable Tree	
Newick Ladderized Tree	
Newick Tree	
Pairwise Distance Table	

D - Virulence Analysis

Reporting unexpected virulence or antibiotic-resistance genes

D1 - Virulence Analysis	
D.1.1 - Virulence Gene Category	**Total Count**
antibiotic resistance	23

transcription factor	3
virulence protein	276

D2 - Virulence Analysis

D.2.1 - Highest Ranked Sequences by Virulence Similarity

Matched 76 / 93 total francisella virulence genes

Sequence/Contig Name	Length(bp)	Ref Gene Count	Shared	Query Missing	Ref Missing
MR_4338001 Francisella tularensis subsp tularensis FTI - SFG - WyNC Francisella	1,898,494	88	73	15	3
MR_4338201 Francisella tularensis FTN - SFG - FRAN047 type A Glued fragments of sequence 622147 Francisella tularensis FTN FRAN047 draft sequence	1,848,610	88	73	15	3
MR_4342301 Francisella tularensis subsp tularensis WY96-3418 Francisella	1,898,476	88	73	15	3
MR_4342201 Francisella tularensis subsp tularensis FSC033 Glued fragments of sequence 695074 Francisella tularensis subsp tularensis FSC033 Francisella tularensis subsp tularensis FSC033 unfinished sequence whole genome shotgun sequencing project	1,847,789	89	73	16	3
MR_4342101 Francisella tularensis subsp tularensis FSC 198 Francisella	1,892,616	89	73	16	3
MR_4340001 Francisella tularensis subsp novicida GA99-3548 Glued fragments of sequence 695069 Francisella tularensis subsp novicida GA99-3548 Francisella tularensis subsp novicida GA99-3548 unfinished sequence whole genome shotgun sequencing project	1,849,843	87	72	15	4
MR_4342401 Francisella tularensis subsp tularensis FTB - SFG - UNMC 061598 Francisella	1,892,681	89	73	16	3
MR_4342001 Francisella tularensis subsp tularensis SCHU S4 Francisella	1,892,819	89	73	16	3
MR_4338301 Francisella tularensis subsp UNKNOWN FTJ - SFG Wy4414 Glued fragments of sequence 657722 Francisella tularensis subsp UNKNOWN FTJ draft sequence	1,863,834	87	72	15	4
MR_4340501 Francisella novicida 11B Glued fragments of sequence 706943 Francisella novicida 11B	1,873,974	86	71	15	5

draft sequence					

D.2.2 - Virulence Gene Category	Count
antibiotic resistance	23
transcription factor	3
virulence protein	271

D.2.3 - Virulence Significance					
Top Species	Sequence Count	Gene Count	Shared	Query Missing	Species Missing
Francisella tularensis	22	64	54	17	2

D.2.4 - Phylogenetic Tree	
Sequence Group Count	43
Sequence Group Type	all sequence
Pairwise Gene Comparison	symmetric difference
Phylogenetic Tree Pdf	
Newick Readable Tree	
Newick Ladderized Tree	
Newick Tree	
Pairwise Distance Table	

E - G/E Vector Analysis

Finding evidence of bacterial vector-mediated genetic engineering

E.1 - Genetic Engineering
No G/E Results

E.2 - Genetic Engineering - Blast Results (top 10)

GE Vector	Length(bp)	CoverageLength(bp)	Coverage %
gbdn_gi|1684862|gb|U72488.1|CVU72488 Cloning vector pRNA8, complete sequence	11,918	4,575	38.39
pKK3535.Generic.html	11,796	4,574	38.78
gbdn_gi|559545|gb|U12809.1|XXU12809 Transformation vector pPRV1, plastid targeting seqment	2,962	1,313	44.36
gbdn_gi|559547|gb|U12811.1|XXU12811 Transformation vector pPRV100B, plastid targeting segment	3,019	1,313	43.52

gbdn_gi|559546|gb|U12810.1|XXU12810 Transformation vector pPRV100A, plastid targeting segment	3,019	1,313	43.52
gbdn_gi|559552|gb|U12814.1|XXU12814 Transformation vector pPRV112A, plastid targeting segment	4,126	1,314	31.85
gbdn_gi|559554|gb|U12815.1|XXU12815 Transformation vector pPRV112B, plastid targeting segment	4,126	1,314	31.85
gbdn_gi|80261323|gb|DQ211347.1| Plastid transformation vector pPRV110L, complete sequence	4,128	1,313	31.83
gbdn_gi|559550|gb|U12813.1|XXU12813 Transformation vector pPRV111B, plastid targeting segment	4,174	1,313	31.48
gbdn_gi|559548|gb|U12812.1|XXU12812 Transformation vector pPRV111A, plastid targeting segment	4,174	1,313	31.48

Software Packages List

We utilize the following software components in our system. Unless otherwise noted, all are freely available (Open Source) software.

Modules

- BLAST-2.21
- MUMmer 3.22
- mysql-5.1.36
- Oracle (Note: LLNL has a site license. Oracle is not freely available. However, the free MySQL database could be readily substituted as we have not utilized any Oracle-specific features.)
- phylip-3.69
- Python-2.6.2
    - amqplib-0.6.1
    - anyjson-0.2.5
    - biopython-1.53
    - carrot-0.10.7
    - celery-2.1.1
    - cx_Oracle-5.0.2
    - distribute-0.6.13
    - django
    - django-celery-2.1.1
    - django_debug_toolbar-0.8.3
    - django_pagination-1.0.7
    - django_picklefield-0.1.6
    - django_piston-0.2.2
    - docutils-0.6
    - ete2-2.0rev111
    - httplib2-0.6.0
    - Imaging-1.1.7
    - importlib-1.0.2
    - matplotlib-1.0.0
    - MySQL-python-1.2.3c1
    - networkx-1.3
    - newick-1.3
    - nose-0.11.3
    - numpy-1.3.0
    - paramiko-1.7.4
    - pip-0.7.2
    - psycopg2-2.0.14
    - pycrypto-2.0.1
    - pyparsing-1.5.5
    - PyGreSQL-4.0
    - python_dateutil-1.5
    - PythonQt2.0.1

    - PyYAML-3.09
    - ReportLab_2_4
    - scipy-0.7.1
    - setuptools-0.6c11
    - SQLAlchemy-0.6.5
    - yolk-0.4.1
- rabbitmq_server-2.1.0
- R-2.9.2
    - RColorBrewer_1.0-2
    - Plotrix_2.7-2

TcPipeline

- __init__.py
- BasePathway.py
- FileManager.py
- forms.py
- GeneDataSetup.py
- GeneFamilyPathway.py
- GeneSummary.py
- geneViews.py
- getParentInfo.py
- GetSequenceInfo.py
- MfeBatchProcessTritool.py
- MfeMailWorker.py
- MikiTest.py
- model_utils.py
- models.py
- MummerBatchPreparer.py
- MummerManager.py
- notifymfe.py
- NucmerReportParser.py
- PathwayDirector.py
- sshupload.py
- TaskManager.py
- tasks.py
- tc_utils.py
- tcpipeline_config.py
- tests.py
- tritool_config.py
- tritool_orm.py
- TritoolBasePathway.py
- TritoolGePathway.py
- TritoolMailParser.py
- TritoolRequester.py

- TritoolSdPathway.py
- TritoolSequence.py
- TritoolSnpPathway.py
- TritoolSnpPathwayFactory.py
- urls.py
- views.py
- VirulencePathway.py

MFE Tools

- __init__.py
- demoMummerBatchPreparer.py
- MfeBlastRunner.py
- MfeFastaFileLoader.py
- MfeGeneLoader.py
- MfeNdf.py
- MfeNdfDriver.py
- MfeNdfLoader.py
- MfeParentLoader.py
- MfeSnpLoader.py
- MfeVectorProbeLoader.py
- models.py
- MummerBatchPreparer.py
- tests.py
- urls.py
- views.py

MFE sigFinder

- __init__.py
- admin.py
- forms.py
- inputsequences.py
- models.py
- subtree.py
- taskgroup.py
- tests.py
- urls.py
- views.py

Tri-Tool psi-kit

- analyze_genotypes.py
- assay_config.py
- extract_snps.py
- fmt_expt_data.py
- gene_probes.py

- geno_app_files.py
- init_genome_data.py
- make_vmatch_tag.py
- newick_format.py
- newick_ladderize.py
- newick_treeorder.py
- newick2pdf.py
- parsimony_treebuilder.py
- prep_all.py
- psi_analyze.py
- psi_prep.py
- psikit_show_assay_genomes.py
- req_set_error_msg.py
- show_assay_genomes.py
- vmatch_alleles.py

Tri-Tool detection

- blast_hit_fmdv.py
- blast_hit.py
- combine_blastdb.py
- createParams.py
- det_evidence.py
- generic_targ.py
- init_model_params.py
- kpath_org.py
- mle_analyze_heapy.py
- mle_analyze_mod.py
- mle_analyze.py
- object_persist.py
- parse_blast.py
- probe.py
- project.py
- sim_detect.py
- submit_format_blast_results.py
- submit_my_pblast.py
- targ_evidence.py
- target_taxonomy.py
- target.py